Enhancer Prediction in Proboscis Monkey Genome: A Comparative Study

Norshafarina Omar[1], Yu Shiong, Wong[1], Xi, Li[2] and Yee Ling, Chong[3], Mohd Tajuddin Abdullah[4] and Nung Kion, Lee[1]
[1]*Department of Cognitive Sciences, Universiti Malaysia Sarawak.*
[2]*Life Science Informatics, Data 61, CSIRO.*
[3]*Faculty of Resource Sciences and Technology , Universiti Malaysia Sarawak.*
[4]*Kenyir Research Institute, Universiti Malaysia Terengganu.*
nklee@unimas.my

***Abstract*—Genome annotation is an essential task for understanding and analyzing the whole genome and its function. We have sequenced the complete proboscis Monkey (*Nasalis larvatus*) genome due to its importance for medical and evolutionary studies. We have performed an initial annotation of the genes genome using the MAKER gene annotation pipeline. 3084 genes were predicted from chromosome 18 of the genome using six eukaryotic model species. Intergenic regions possibly enriched with enhancers are then predicted using five different tools: DeepBind, LS-GKM, GMFR-CNN, CSI-ANN and iEnhancer-2L. These tools find the enhancers of the complex intergenic regions based on epigenetic features, in which intergenic regions are seen as a potential region for enhancers with a certain epigenetic features bound to it. Empirical results demonstrate competitive performance using different prediction tools with multiple epigenetic features to predict the enhancers for chromosome 18 in *proboscis monkey*. Based on the findings of this study, predicted enhancers can be used for the purpose of scientific and genomic discoveries.**

***Index Terms*—Enhancer Annotation; Enhancer Prediction; Motif Discovery; Proboscis Monkey.**

I. Introduction

Annotation is the first step in understanding the biological functions, identifying functional elements, and for performing scientific inquiries using the genome of a species. Fundamental annotation tasks includes identifying coding and non coding DNA regions in a genome. Regulatory elements are important functional DNA sequences located in non coding region of a genome. They play a major role in regulating gene expression for the production of RNA and proteins. Regulatory elements include promoters, enhancers, proximal regulatory and distal regulatory elements. Predicting enhancer is one of the important tasks since enhancer has a capability to regulate gene expression. However, experimental approaches are costly and time consuming, therefore, a reliable and effective computational approach is needed for annotation of enhancers.

There were several studies of experimental approaches and computational approaches which have been done with enhancer prediction. Liu et al. [1] aimed to identify enhancers along with their strength by using the pseudo *k*-tuple nucelotide composition in order to formulate the DNA sequences. Meanwhile, Dai et al. [2] investigate the relationship between low methylated regions (LMRs) that derived from whole genome bisulfite sequencing (WGBS) with the enhancer prediction. Some studies learned enhancers from DNA sequence features by capturing the combination of binding sites [3]. Since enhancer tend to be bound on certain epigenetic features, [4,5,6] combined transcription factors, and chromatin histone modifications to identify enhancers and it has been found to improve the accuracy of enhancer predictions. According to Zhu et al. [7] enhancers are generalized as the peaks of the H3K4me1 enriched regions, and it was supported by [6,8] where the presence of this histone modification along with H3K4 methylation, H3K27ac and few transcription factors (TF) such as EP300, CTCF, TAL1, GAT1 were used to predict the enhancers [5].

Different enhancer prediction tools have been developed and widely used. [1] used SVM to distinguish enhancers from the whole genome sequences. In [9], they proposed an enhancer predictor called DELTA by integrating shape features of histone modifications with AdaBoost algorithm. The DeepBind [10] is one of the promising pattern discovery, a tool that is based on deep convolutional neural networks. [6] identified functional DNA features by making use of chromatin signatures and applied artificial neural network on it. Wong et al. [11] proposed an integrated enhancer predictor based on gapped motif features representation (GMFR) and deep convolutional neural network (CNN).

II. Related Work

MAKER is an automated gene annotation pipeline that mainly include masking repetitive elements, *ab initio* gene prediction using programs such as: SNAP, AUGUSTUS and GeneMark, aligning the predicted *ab initio* gene models together with reference protein sequences and transcript sequence (EST/RNA) from closely related species, applying certain refinement metrics to produce the final annotated gene models. For more details about MAKER pipeline and how it works, readers can refer to refs [12,13]. Coombe et al. [14] used MAKER to annotate the coding and non-coding genes of Sitka spruce and used gene sequences of Norway spruce as evidences. MAKER also has been used to annotate the whole genome of desert woodrat [15]. A total of 24,574 coding genes

were annotated in desert woodrat genome using two gene prediction programs; SNAP and Augustus with evidences from mouse and rat proteins. In [16], the authors annotated 61,773 genes from valley oak genome using seven plant species as their evidences and three gene predictors including SNAP, Augustus, and FGENESH. By annotating the brewer's yeast genome using one reference for both transcript and protein sequence evidence, MAKER was able to annotate 9,939 genes [17]. Another genome annotation using MAKER was done by Choo et al. [18] to annotate pangolin genomes. They used multiple evidences to predict 23,446 and 20,298 genes in the two pangolin species, which is based on *ab initio* gene prediction, transcriptomic data and protein evidence from one different species as reference genome.

III. Genome Annotation

The enhancer annotation for chromosome 18 in proboscis monkey consists of several steps (Figure 1). One of our goals is to identify the gene regions within the chromosome 18. Chromosome 18 of *proboscis monkey* can be accessed at NCBI GenBank under accession number GCA_000772465.1. We used MAKER annotation pipeline to annotate our genome and as for reference data, we collected the annotated protein and transcript sequences from six different species which are *gorilla gorilla*, *macaca mulatta*, *mus musculus*, *pan troglodytes*, *homo sapiens* and *pongo abelii* (ftp://ftp.ncbi.nlm.nih.gov/genomes/). In addition, we used three gene prediction programs: SNAP, Augustus, and Genemark. In total, we have identified 3084 genes in chromosome 18 of proboscis monkey. During the annotation process, we first masked repetitive elements from the Chromosome 18, then carried out the gene prediction and aligned the evidence protein and EST sequences from six species using BLAST. We further refined the *ab initio* gene models predicted by SNAP, Augustus and GeneMark together with the aligned evidences through MAKER pipeline. After that, we proceed with the extraction process to identify the intergenic regions, which possibly contain enhancers. In order to extract the intergenic regions, we used GFF-Ex [19] to process the gff file generated from MAKER. GFF-Ex is able to extract sequences based on numerous region boundaries such as exons and introns regions, gene regions, upstream regions and also the intergenic regions. GFF-Ex extracted 1783 intergenic regions, 19 782 exon and 15 769 introns. The final was constructed by excluding intergenic regions with length less than 500bp and only considered those located 500bp away from the transcription start site (TSS) and transcription end site (TSE).

Table 1
Summarized description of benchmark datasets

Tool	Benchmark Dataset
DeepBind	CTCF,EP300
LS-GKM	CTCF,EP300
GMFR-CNN	CTCF,EP300
CSI-ANN	H3K4me1
iEnhancer-2L	H3K4me1, H3K4me3,H3K27ac,etc.

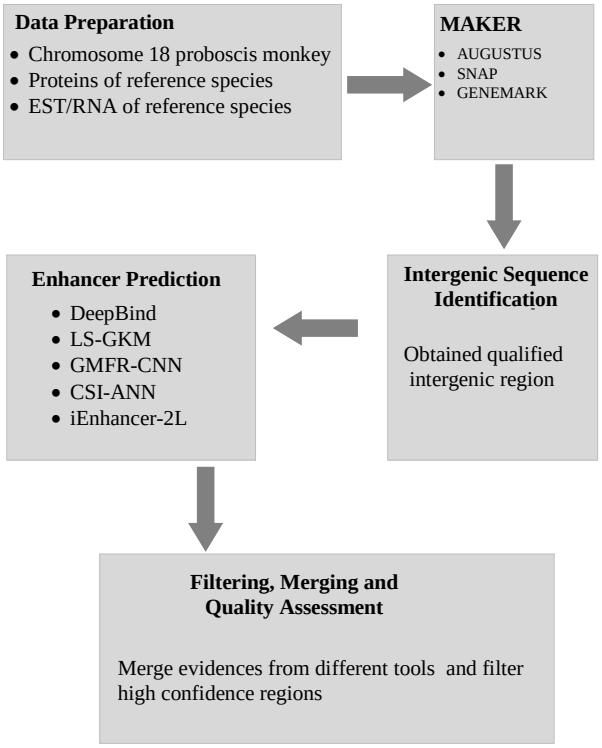

Figure 1: Enhancer Annotation Pipeline

IV. Enhancer Prediction in Proboscis Monkey

To predict enhancers, we employed five computational tools including DeepBind [11], LS-GKM [20], GMFR-CNN [12], CSI-ANN[6] and iEnhancer-2L[1]. DeepBind is a deep convolutionary neural networks that learn to model motifs in datasets using one-hot encoding of input DNA sequences. LS-GKM is based on creating a prediction model using SVM with k-mer feature as inputs. CSI-ANN is based on fisher discriminant analysis and time delay neural network that learn the features using chromatin signals. GMFR-CNN is a convolutionary neural networks based on the dependencies feature in the k-mers. Meanwhile iEnhancer is an SVM predictor based on the feature in the k-tuple nucleotide. Table 1 summarized the datasets used by each of the tool for prediction of enhancers. For DeepBind, LS-GKM, and GMFR-CNN, we used the binding sequences of CTCF transcription factor and sequences associated with co-factor EP300. Both datasets are known to be associated with enhancers [5,21]. Since not all enhancers are associated with those two datasets, additional two tools CSI-ANN and iEnhancer-2L which utilitized histone datasets are employed. CSI-ANN used H3K4me1 histone marks for predicting enhancer, meanwhile iEnhancer-2L is based on multiple histone marks including H3K4me1, H3K4me3,H3K27ac, etc.

The input to the computational tools were obtained via the previous annotation. There are a total of 1783 intergenic sequences are extracted from Chromosome 18. Next we input the intergenic sequences to DeepBind, LS-GKM, GMFR-CNN, CSI-ANN and iEnhancer-2L for enhancer locations prediction. However, there are two primary challenges to predict enhancers using different prediction tools. First, each tools may have required different input features. Second, the parameters for

each tools are different from one another. Some tools may have few parameters that need to be tuned, but some may have numerous of it. The parameters used for each tools are summarized in Table 2.

Table 2
Parameters used by different prediction tools

Tools	Parameter
DeepBind	We used trained model provided by DeepBind; Consist of input layer,convolutional layer,rectification layer, pooling layer, neural network prediction layer and an output layer; Learning rate=0.0005,0.05; Learning momentum=0.95,0.99 Number of iteration=4000-20 000
LS-GKM	We trained the model using CTCF and EP300 datasets; l =14(CTCF), 9(EP300); k =6 (CTCF), 6(EP300)
GMFR-CNN	Consist of 6 layer (input layer,2 concolutional layers,2 subsampling layers,an output layer); We trained the model using CTCF and EP300 datasets; Learning rate=0.8; Number of iteration =200
CSI-ANN	Time delay neural network(TDNN) classifier with a delay of 9,2 hidden layer nodes and an output layer; We treained the model using chromosome18 intergenic regions;
iEnhancer-2L	w is train using particle swarm optimization (PSO) 2-layer predictor with 2968 trained samples obtained from final benchmark daraset; k = 6 (1st and 2nd layers); w = 0.1(1st layer), 0.4(2nd layer); λ =9 (1st and 2nd layers)

Using two transcription factors, both DeepBind, LS-GKM and GMFR-CNN produced two set of predicted enhancers. We combined those two sets of predicted enhancers from each tool into a single file. We then merged these files using bedtools merge utility[22]. The *merge* utility allowed us to merge features that are overlapping. In Figure 2, we show an example of merging two overlapped predicted enhancer regions into single region using *bedtools merge*.

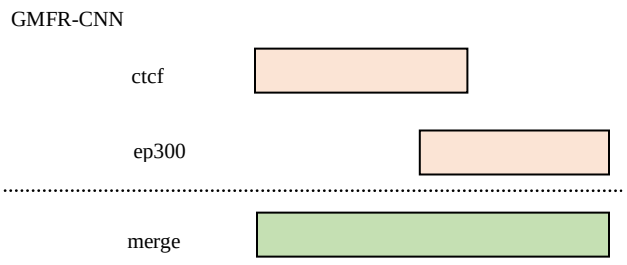

Figure 2: Merging predicted enhancer regions

The predicted enhancers for chromosome 18 are chosen based on overlapping features obtained by all five tools. In order to determine if any of the features in the two, three, four or five tools are overlapping with one another, we used *bedtools intersect*. Figure 3 shows an example of using bedtools function called intersect. This *bedtools intersect* aimed to identify any common features between two or more set of genomic features. Now using the same *bedtools merge*, we merged the overlapping features by varying the combination of tools, as will be further discussed on the next section.

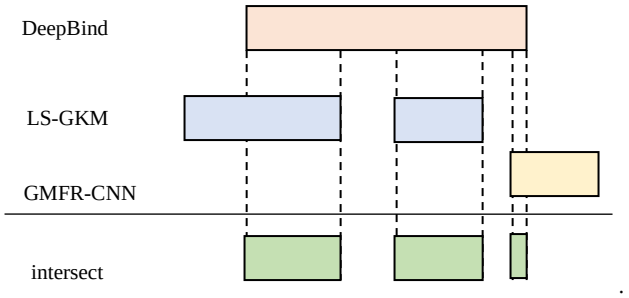

Figure 3: Intersection between tools

V. Results and Discussions

In order to evaluate the performance of the predictor tools in predicting enhancers, we used intergenic sequences generated from the previous annotation process and we limited the analysis to chromosome 18 in proboscis monkey. In Table 3, we listed the number of predicted enhancers of DeepBind, LS-GKM, GMFR-CNN, CSI-ANN and iEnhancer-2L. The number of predicted enhancers for CSI-ANN and iEnhancer-2L are 22,954 and 14,700 respectively. As mentioned in Section IV, by using the *bedtools merge* for DeepBind, LS-GKM and GMFR-CNN, we combined the overlapping features for DeepBind, LS-GKM and GMFR-CNN and thus we obtained 31,805, 65,349 and 73,133 enhancer regions, respectively.

Table 3
Performance of Each Tools

Tool	Number of Predicted Enhancer
DeepBind	31 805
LS-GKM	65 349
GMFR-CNN	73 133
CSI-ANN	22 954
iEnhancer-2L	14 700

In addition, we computed the coverage values for each tools by using bedtools functions called coverage. This *bedtools coverage* is used to find the coverage of a single tool features compared to the other four tools features coverage (Table 4) and to count the mapped reads on the predicted enhancer regions and on the non-enhancer regions of chromosome 18 in *proboscis monkey* genome. To compute the percentage of coverage, we calculated the sum of a fraction of bases in each tools that had coverage in four other tools. And then we divided by the total predicted enhancers to obtain the percentage of coverage. We performed coverage analysis because we wanted to measure the sensitivity of the tools. The sensitivity of the tools which based on the coverage across the predicted enhancer regions are reported in Table 4. The number of predicted enhancers features in GMFR-CNN tool had highest coverage (79.04%) from features in GMFR-CNN, LS-GKM, CSI-ANN and iEnhancer-2L, followed by DeepBind with 37.90% of coverage from features in four other tools. LS-GKM and iEnhancer-2L almost had similar coverage, 23.04% and 28.70% respectively, with iEnhancer-2L had slighty higher

percentage of coverage than LS-GKM. On the other hand, CSI-ANN had the less coverage among all the tools (15.61%). The overall mean sensitivity of the tools is 36.86%.

Table 4
Coverage Values for Each Tools

Tool	Sum	Total Enhancer Features	Percentage of Coverage (%)-Sensitivity
DeepBind	12053.15	31 805	37.90
LS-GKM	15055.69	65 349	23.04
GMFR-CNN	57803.93	73 133	79.04
CSI-ANN	3582.889	22 954	15.61
iEnhancer-2L	4218.954	14 700	28.70

We have performed many experiments varying combination of different tools among those five listed tools. For example using two tools, we run 20 combination of tools. In three tools, the experiments can be combined in 60 ways. There are 120 and 20 ways of combination using four and five tools respectively, to identify the overlapping predicted enhancer regions. Table 5 listed the number of predicted enhancer for 20 combination of two tools. Using all the result from these combination, we combined it and removed the redundance predicted enhancers to finalize the exact number of predicted enhancers.

Table 5
Numbers of overlapped enhancers predicted by two tools

	DeepBind	LS-GKM	GMFR-CNN	CSI-ANN	iEnhancer-2L
DeepBind	-	13872	28874	1707	9079
LS-GKM	13872	-	57695	3513	21208
GMFR-CNN	28874	57695	-	3861	22341
CSI-ANN	1707	3513	3861	-	3394
iEnhancer-2L	9079	21208	22341	3394	-

To compare and contrast different combinations of tools, we used bedtools intersect to find the overlapping features between combination of any two tools. As shown in Table 6, using 3 different tools, we can see that the number of overlapping features among those number of tools is slightly higher compared to others. Reading from Table 5, using two and five different tools has less in number of overlapping enhancers. Probably, this is due to the less number of combination tools and as a result, not many overlapped are found between the overlapping features from two and five tools.

Table 6
Number of overlapping enhancers

Number of tools	Number of overlapping enhancer features
2	165544
3	500145
4	496632
5	3861

Because the combination of tools (from 2,3,4 and 5 tools) may predict the same enhancer location, we finalized the predicted enhancers by merging the results using the *bedtools merge*. In Table 7, it showed that enhancer features extracted by using 2, 3 and 4 number of tools generated the same number of predicted enhancers. Not just the number of predictions are the same, but they did predict the same location of enhancers on chromosome 18 of proboscis monkey. Although these tools using different epigenetic features as benchmark to extract enhancer location in proboscis genome, but they did predict the same enhancer features. This suggest that using multiple epigenetic features to predict enhancer might improve the prediction. The fact that enhancer is not only occupied by certain transcription factors but also different histone marks.

Table 7
Number of predicted enhancers

Number of tools	Number of predicted enhancers
2	77387
3	77387
4	77387
5	3861

VI. CONCLUSION

The goal of this study is to identify the potential enhancer regions in *proboscis monkey* for the purpose of scientific and genomic discoveries. By using different epigenetic features and enhancer associated TFs and co-factor to predict the enhancers, we have achieved promising results. The combination of different enhancer prediction tools along with multiple epigenic features is capable of predicting almost similar enhancer features. Other epigenic features that have not been included in this study such as DNase I hypersensitivity, TAL1 and GATA1 can also be included, this should improve the predicting of the enhancer. However, the limitation of this study is that the annotation process required long computational runtime. Further study may focus on other chromosomes in *proboscis monkey* genome.

ACKNOWLEDGMENT

This study is funded by the Ministry of Higher Education through the Fundamental Research Grant Scheme.

REFERENCES

[1] B. Liu, L. Fang, R. Long, X. Lan and K.-C. Chou, "iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition," *Bioinformatics,* vol. 32, pp. 362-369, 2016.

[2] Y. Dai, J. Xu and H. Hu, "LMethyR-SVM: Predict human enhancers using low methylated regions based on weighted support vector machines," *bioRxiv,* p. 054221, 2016.

[3] D. Lee, R. Karchin and M. A. Beer,"Discriminative prediction of mammalian enhancers from DNA sequence," *Genome research,* vol. 21, pp. 2167-2180, 2011

[4] M. Fernández and D. Miranda-Saavedra,"Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines," *Nucleic acids research,* vol. 40, pp. e77-e77, 2012.

[5] N. Dogan, W. Wu, C. S. Morrissey, K.-B. Chen, A. Stonestrom, M. Long, C. A. Keller, Y. Cheng, D. Jain and A. Visel, "Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility," *Epigenetics & chromatin,* vol. 8, p. 1, 2015.

[6] H. A. Firpi, D. Ucar and K. Tan, "Discover regulatory DNA elements using chromatin signatures and artificial neural network," *Bioinformatics,* vol. 26, pp. 1579-1586, 2010.

[7] Zhu, Y., Sun, L., Chen, Z., Whitaker, J. W., Wang, T., & Wang, W. (2013). Predicting enhancer transcription and activity from chromatin modifications. *Nucleic acids research, 41*(22), 10032-10043.

[8] B. E. Blaisdell, "A measure of the similarity of sets of sequences not requiring sequence alignment," *Proceedings of the National Academy of Sciences*, vol. 83, pp. 5155-5159, 1986.

[9] Y. Lu, W. Qu, G. Shan and C. Zhang, "DELTA: a distal enhancer locating tool based on AdaBoost algorithm and shape features of chromatin modifications," *PLoS ONE*, vol. 10, p. e0130622, 2015.

[10] B. Alipanahi, A. Delong, M. T. Weirauch and B. J. Frey, "Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning," *Nature biotechnology*, 2015.

[11] Y.S.Wong, N.K.Lee, N.Omar, "GMFR-CNN:an integration of gapped motif feature representation and deep learning approach for enhancer prediction*,"7th International Conference on Computational Systems-Biology and Bioinformatics*, 2016 (In press)

[12] M. S. Campbell, C. Holt, B. Moore and M. Yandel, "Genome annotation and curation using MAKER and MAKER-P," *Current Protocols in Bioinformatics*, pp. 4.11. 1-4.11. 39

[13] B. L. Cantarel, I. Korf, S. M. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A. S. Alvarado and M. Yandell, "MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes," *Genome research*, vol. 18, pp. 188-196, 2008.

[14] L. Coombe, R. L. Warren, S. D. Jackman, C. Yang, B. P. Vandervalk, R. A. Moore, S. Pleasance, R. J. Coope, J. Bohlmann and R. A. Holt, "Assembly of the Complete Sitka Spruce Chloroplast Genome Using 10X Genomics' GemCode Sequencing Data," *PLoS ONE*, vol. 11, p. e0163059, 2016

[15] M. Campbell, K. F. Oakeson, M. Yandell, J. R. Halpert and D. Dearing, "The draft genome sequence and annotation of the desert woodrat Neotoma lepida," *Genomics Data*, vol. 9, pp. 58-59, 2016.

[16] V. L. Sork, S. T. Fitz-Gibbon, D. Puiu, M. Crepeau, P. F. Gugger, R. Sherman, K. Stevens, C. H. Langley, M. Pellegrini and S. L. Salzberg,"First Draft Assembly and Annotation of the Genome of a California Endemic Oak Quercus lobata Née (Fagaceae)," *G3: Genes| Genomes| Genetics*, vol. 6, pp. 3485-3495, 2016.

[17] P. M. De León-Medina, R. Elizondo-González, L. C. Damas-Buenrostro, J.-M. Geertman, M. Van den Broek, L. J. Galán-Wong, R. Ortiz-López and B. Pereyra-Alférez, "Genome annotation of a Saccharomyces sp. lager brewer's yeast," *Genomics Data*, vol. 9, pp. 25-29, 2016.

[18] S. W. Choo, M. Rayko, T. K. Tan, R. Hari, A. Komissarov, W. Y. Wee, A. A. Yurchenko, S. Kliver, G. Tamazian and A. Antunes, "Pangolin genomes and the evolution of mammalian scales and immunity," *Genome research*, vol. 26, pp. 1312-1322, 2016.

[19] D. Gupta, "GFF-Ex: A genome feature extraction package," *Journal of Natural Science, Biology and Medicine*, vol. 2, p. 90, 2011.

[20] M. Ghandi, D. Lee, M. Mohammad-Noori and M. A. Beer, "Enhanced regulatory sequence prediction using gapped k-mer features," *PLoS Computational Biology*, vol. 10, p. e1003711, 2014.

[21] S. J. B. Holwerda and W. de Laat, "CTCF: the protein, the binding partners, the binding sites and their chromatin loops," *Phil. Trans. R. Soc. B*, vol. 368, p. 20120369, 2013.

[22] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, pp. 841-842, 2010.